Bayesian Model Averaging

SIBGHAT & SIDDHANT

Background Problem

- Model Uncertainty
- Model Selection
- Better Prediction

BMA Introduction: Notations

- Δ is quantity of interest
- Future Observable
- Utility of a course of action
- D is data
- ***M*** **=** {Mk, k=1,2,3,…,K}
- Θk vector of parameters in Mk
- Pr(θk | Mk) prior density of θk under Mk
- Pr(D | θk, Mk) is likelihood for data
- Pr(Mk) is prior probability that Mk is the true Model

Mathematics for BMA

- Posterior distribution given data D is

$$\Pr(\Delta \mid D)=\sum_{k=1}^{K} pr(\Delta \mid Mk, D)\, pr(Mk \mid D)$$

Posterior probability for model Mk ∈ $\boldsymbol{M}$ is

$$\Pr(\mathrm{Mk} \mid \mathrm{D}) = \frac{\Pr(D \mid Mk)\, pr(Mk)}{\sum_{l=1}^{K} pr(D \mid Ml) pr(Ml)}$$

where

$$\boldsymbol{Pr(D \mid Mk)} = \int \boldsymbol{Pr(D} \mid \theta k, \boldsymbol{Mk)}\, \boldsymbol{pr(}\theta k \mid \boldsymbol{Mk)}\, \boldsymbol{d}\theta k$$

Expectation using BMA

- Let Δˆk = E [Δ | D, Mk]

- Posterior mean of Δ using BMA is

$$E[\Delta \mid D] = \sum_{k=1}^{K} \hat{\Delta}_k \, pr\,(Mk \mid D)$$

Implementation Issues

- Simple right ?

- Not so Much….

Issues …

- BMA is nice but difficult to implement

- Reasons:
- ***M*** can be enormous. Infeasible to sum over all models
- Integrals can be hard to compute, even using MCMC
- What about Prior distribution ?
- Is it worth it ?

Managing the Summation

- Feasible way to compute the equation

$$\sum_{k=1}^{K} \Pr(\Delta \mid Mk, D) \Pr(Mk \mid D)$$

Two approaches

- Occam's window method
- Monte Carlo Markov Chains Model Composition

Occam’s window method

- Average over a subset of models supported by the data

- Principle 1
- Disregard a model if it predicts the data far less than the model with best predictions
- Formally:

$$A' = \{Mk;\ \frac{\max l\ (\Pr(Ml\ |D))}{\Pr(Mk\ |D)} \leq C\}$$

Occam's Window method …

- Exclude complex model if the data supports the simpler model (Occam's razor)
- Formally

$$B=\{Mk: \exists\ Ml \in A', Ml \subset Mk, \Pr(Ml \mid D) / \Pr(Mk \mid D) > 1\}$$

Subset of models to be used is : A = A' \ B

All probabilities conditional on A

Use MCMC Model composition

- Use MCMC to directly approximate the first equation

- Construct a Markov Chain to {M(t)}, t=1,2,… with state space ***M*** and stationary distribution Pr(Mi | D)

- Simulate chain to get observations M(1),…..,M(N)
- Then for any function g(Mi) defined on ***M***, compute average

- $$\boldsymbol{G(est)} = \frac{\sum_{t=1}^{N} g(M(t))}{N}$$

Computing Integrals

- Integrals of the form:

$$\boldsymbol{Pr}(\boldsymbol{D}\ |\boldsymbol{Mk}) = \int \boldsymbol{Pr}(\boldsymbol{D}\ |\theta\text{k}, \boldsymbol{Mk})\ \boldsymbol{pr}(\theta\text{k}\ |\boldsymbol{Mk})\ \boldsymbol{d}\theta\text{k}$$

Can be difficult to compute.

Solution ?

Computing Integrals…

- Closed form integrals available for multiple Regression & graphical models

- Laplace method helps approximate Pr(D | Mk) & sometimes yields BIC approximation

- Approximate Pr(Δ | Mk, D) with Pr(Δ | Mk, θ(estimated), D) where θ(estimated) is MLE

BMA for Linear Regression: Predictors, Outliers & Transformations

- Suppose a dependent variable Y and predicts X1,…..Xk. Then variable selection methods try to find the “best” model with the form

$$\mathrm{Y} = \sum_{j=1}^{p} Bij\ Xij + \varepsilon + \mathrm{B0}$$

BMA however tries to average over all possible sets of predictors

Linear Regression: Transformation & Outliers

- We can use Box-cox transformation for the response

$$y^{(p)}=\begin{cases} \frac{y^p-1}{p} \; if\ p\ not\ equal\ to\ zero \\ \log(y)\ if\ p\ equal\ to\ zero \end{cases}$$

And the model is $y^{(p)}$=XB+ε where ε~N(0,σ^2I)

We can use "change point transformations" to transform the predictors

- Use the output from the alternating conditioning expectation algorithm to suggest the form of transformation
- Use Bayes factor to choose the precise transformation

Linear Regression: Transformation & Outliers

- We can use variance-inflation model for outliers by assuming:

$$\varepsilon = \begin{cases} N(0, \sigma^2)\ w.p & (1-\pi) \\ N(0, K^2\sigma^2) & w.p\ \pi \end{cases}$$

Simultaneous variable & Outlier selection method

- Use a highly robust technique to identify potential outliers
- Compute all possible posterior model probabilities or use MC3, considering all possible subsets of potential outliers.

Generalized Linear Models

- The Bayes factor for model M1 against M0:

$$B10 = \Pr(D \mid M1) \big/ \Pr(D \mid M0)$$

Consider M+1 models M0,M1,……,Mk. Then the posterior probability for Model Mi using Bayes Factor is:

$$\Pr(Mi \mid D) = \frac{\alpha_{iB_{i0}}}{\sum_{j=0}^{K} \alpha_{jB_{j0}}}$$

Where $\alpha_{i} = \frac{\Pr(Mi)}{\Pr(M0)} \quad i = 0, \ldots, K$

Generalized Linear Model

- Dependent variable: Y_i

- Independent variables: X_i=(x_{i1}, ….., x_{ip}), i=0,….,n where x_{i1}=1

- The null model M0 is defined as B_j=0 where j=2,…..,p

Generalized Linear Models ....

- If we use Laplace approximation, we get

$$\Pr(D \mid Mk) \approx (2\pi)^{p_k/2} \mid \Psi \mid^{1/2} \Pr(D \mid \hat{B_k}, Mk) \Pr\left(\hat{B_k} \middle| Mk\right)$$

Where p_k is the dimension of $B_{k,}$ and $\hat{B_k}$ is the posterior mode of $B_{k,}$ and Ψ_k is minus the inverse hessian of

$h(B_{k,}) = \{\Pr(D \mid B_{k,}Mk) \Pr(B_{k,} \mid Mk)\}$ evaluated at $B_{k,} = \hat{B_k}$

Generalized Linear Models …

- Suppose E[B_k | Mk] = w_k and var(B_k | Mk) = W_k
- Use one step of Newton's method to approximate $\hat{B_k}$ starting from $\hat{B}$
- Then we have the approximation

$$2 \log B_{10} \approx X^2 + (E1 - E0)$$

And

$$X^2 = \{l1\left(\hat{B_1}\right) - l0(\hat{B_0})\}$$

$$l_k(B_k) = \log\{\Pr(D \mid B_k, Mk)\}$$

$$E_k = 2\lambda_k(\hat{B_k}) + \lambda'_k(\hat{B_k})^T (F_k + G_k)^{-1}\{2 - F_k (F_k + G_k)^{-1}\} \lambda'_k (\hat{B_k}) - \log|F_k + G_k| + p_k \log(2\pi)$$

Where F_k is the expected fisher information matrix, $G_k = (W_k)^{-1}$
and $\lambda_k = logPr(B_k \mid Mk)$

Survival Analysis

- Hazard rate λ(t)=$f(t)/1 - F(t)$
- Cox proportional hazard model:λ(t | Xi) = λ0 (t) exp(Xi B) where λ0 (t) is the baseline hazard rate at time t.
- The estimation of B is based on partial likelihood.

$$\mathrm{PL(B)} = \prod_{i=1}^{n} \left(\frac{\exp(Xi\ B)}{\sum_{l\, \in Ri} \exp({X_l}^T B)} \right)^{w_i}$$

Where Ri is the risk at time t_i and w_i indicates whether subject i is censored or not

Survival Analysis …

- A lot of studies have used the MLE approximation for survival analysis models

$$\Pr(\Delta \mid Mk, D) \approx \Pr(\Delta \mid Mk , \widetilde{B}_k, D)$$

And the Laplace approximation

$$\log \Pr(D \mid M) \approx log Pr(D \mid Mk, \widetilde{B}_k) - d_k \log(\mathrm{n})$$

Where d_k is the dimension of $\widetilde{B}_k$

Graphical Models: Missing Data & Auxiliary Variables

- A graphical model is a statistical model with a set of conditional independence relationships being described by means of a graph.
- We'd study here only acyclic direct graph
- Use either analytical or numerical approximations when we apply BMA and Bayesian graphical models to solve problems with missing data. For example

$$\frac{\Pr(D\,|M0)}{\Pr(D\,|M1)}=\mathrm{E}\left(\frac{\Pr(D,Z\,|M0)}{\Pr(D,Z\,|M1)}\;|D,M1\right)$$

Where Z denotes the missing data and/or auxiliary variables

Specifying prior model probabilities

- It has been proved that in BMA informative prior have better predictive performance than the neutral priors
- Consider the equation

$$\Pr(\mathrm{Mi}) = \pi_j^{\delta_{ij}} (1-\pi_j)^{1-\delta_{ij}}$$

Where Mi is a linear model with p covariates, and π_j is the prior probability that $B_j \neq 0$ and δ_{ij} indicates whether X_j is included in Mi or not

- In the case of graphical models, prior probability for each link is specified and multiplied eventually.

Success of a model/Predictive performance

- Split data in two halves randomly, called training & test data sets

- Predictive log score:

Single Model : $\sum_{d \in test\_data} -logPr(d \mid M, training\_data)$

BMA: $\sum_{d \in test\_data} -\log \sum_{M \in A} \{Pr(d \mid M, training_{data}) \Pr(M \mid training\_data)\}$

Smaller P.L.S indicates better predictive performance

Philosophical debate

- Is it Worth it ?
- Example in R.

Questions ?

- Questions regarding BMA